2650/1G681US1

STRUCTURE-BASED DRUG DESIGN FOR ULP1 PROTEASE SUBSTRATES

This application claims priority under 35 U.S.C. §119 from provisional patent application Serial No. 60/205,336, filed May 18, 2000; which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

The present invention relates to rational drug design of cysteine protease inhibitors based on the crystal structure of a Ulp1 protease trapped with its substrate in a covalent intermediate.

BACKGROUND OF THE INVENTION

Ubiquitin (Ub) and ubiquitin-like (Ubl) proteins modulate protein function in the cell through covalent modification of the target protein. Two proteins that have been identified as belonging to this family are ubiquitin and Small Ubiquitin-related Modifier (SUMO-1). The Ub/Ubl conjugated state of cellular proteins has been associated with several critical pathways, including cellular differation, apoptosis, the cell-cycle, and cellular responses to stress (Hershko and Ciehanover, 1998, Annu. Rev. Biochem., 67:425; Hochstrasser, 1998, Genes and Develop., 12:901; Laney and Hochstrasser, 1999; Cell, 97:427, Saitoh, *et al.*, 1997, TIBS, 22:374; Johnson and Hochstrasser, 1997, Trends. Cell. Biol., 7:408). Alterations in the regulation of the Ub/Ubl pathway has been implicated in several disease states, including tumorogenesis and acute promyelocytic leukemia (Hershko and Ciehanover, 1998, Annu. Rev. Biochem., 67:425; Kamitani *et al.*, 1998, J. Biol. Chem., 273:26675).

Ub and Ubl proteins are covalently attached to their cellular targets via a reversible amide linkage between a lysine ε-amino group on the target and the C-terminus of the Ub/Ubl

protein. The fate of the modified protein depends on if it is conjugated to ubiquitin or SUMO-1. Ubiquinated and poly-ubiquinated proteins are generally targeted to the 26S proteosome for degradation in cell-cycle dependent and cell-cycle independent pathways. Sumoylated protein targets appear to modulate the activity of the targeted system either directly or indirectly by altering cellular localization. Cellular pathways that appear to be modulated by the SUMO pathway include activation of RanGAP1, p53 transcriptional regulation, and lκbα protection from ubiquitination (Matunis, *et al.*, 1996; J. Cell. Biol., 135:1457, Gostissa, *et al.*, 1999, EMBO J., 18:6462; Rodriguez, *et al.*, 1999, EMBO J., 18:6455, Desterro, *et al.*, 1998; Mol. Cell., 2:233). Ub/Ubl modifiers are generally deconjugated from their cellular targets by a class of cysteine proteases, known as deubiquinating (DUB) enzymes.

Studies in yeast show that Smt3, a yeast member of the Ubl protein family, conjugation and deconjugation to proteins appears critical to several yeast functions, including septin ring formation, chromosomal segregation, and progression of the cell cycle. Smt3 shares 47% sequence identity with mammalian SUMO-1 and 17% sequence identity with ubiquitin. Recently, studies showed the identification of a novel *Saccharomyces cerevisiae* gene product termed Ubl-specific protease 1 (Ulp1). In contrast to the ubiquitin DUB system, Ulp1 catalyzes Smt3 processing and Smt3-protein deconjugation.

There is a continuing need in the art to develop methods that enable one to evaluate interactions between proteases and their substrates. The development of an efficient method of trapping the protease and substrate in a complex, which would allow for the study of these interactions is needed. Such a method could allow for the development of compounds that could specifically interact with the amino acids necessary for protease function.

U.S. Patent No. 5,834,228 discloses the use of rational drug design for developing inhibitors of the apopain that will form non-covalent interactions with active site, as based on the crystal structure of the protease and substrate. The reference, however, does not indicate how to trap the protease and substrate in a reaction intermediate state or use this method for identifying compounds that produce covalent interactions with the protease. Thus, there is a need in the art to define active site structures of cystine proteases complexed with their substrates.

SUMMARY OF THE INVENTION

The present invention provides for a composition that is comprised of a polypeptide that comprises the catalytic domain of a SUMO protease and its substrate. In the composition, the SUMO protease is trapped in a deacylation intermediate complex with the substrate.

The invention further provides for a method of forming a complex between a polypeptide that comprises the catalytic domain of a protease with its substrate by (1) combining the protease and substrate in a molar ratio; (2) adding a reducing agent, which is capable of trapping a proteolytic deacylation intermediate complex of the protease and substrate, in an amount that is effective to trap the protease and substrate and (3) adjusting the pH of the mixture to about 7.0.

The present invention provides for polynucleotide sequence that encodes a mutant Ulp1. The mutant Ulp1 may contain an amino acid substitution at position 432, 448, 451 472, 474, 489, 490, 493, or 515. The invention also provides for vectors containing polynucleotide sequences of the mutant Ulp1, cells with the vectors containing the mutant Ulp1 polynucleotide sequences, and the mutant Ulp1 polypeptides.

The present invention also advantageously provides a method of identifying potential substrates of cystine proteases by rational drug design. The method is comprised of designing candidate substrates that form interactions with catalytic amino acids which are identified from computer modeling studies based on the crystal structure of the complex of the protease and substrate.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1A and 1B. Stereo images of the Ulp1 active site in complex with Smt3. Hydrogen bonds are represented as black spheres. Cys580, His541, and Asp351 represent the catalytic triad in the protease fold.

DETAILED DESCRIPTION OF THE INVENTION

The present invention advantageously provides an isolated, preferably purified, trapped protease-substrate complex, preferably a cysteine protease. In particular, protease-

substrate complexes of the invention are those of the sumo pathway, involved in many important cellular processes.

The invention is based, in part, on trapping and crystallization of a Ulp1-Smt3 covalent complex, from which the Ulp1 crystal structure was obtained (coordinates are attached as Table 1 (after the specification and before the claims) and deposited in the Protein Data Bank with accession no. NC-001148). This crystal structure permits the rational drug design of Ulp1 and Ulp1 ortholog inhibitions, which function to promote ubiquitin and ubiquitin-like protein activities in cells.

General Definitions

As used herein, the term "isolated" means that the referenced material is removed from the environment in which it is normally found. Thus, an isolated biological material can be free of cellular components, *i.e.*, components of the cells in which the material is found or produced. In the case of nucleic acid molecules, an isolated nucleic acid includes a PCR product, an isolated mRNA, a cDNA, or a restriction fragment. In another embodiment, an isolated nucleic acid is preferably excised from the chromosome in which it may be found, and more preferably is no longer joined to non-regulatory, non-coding regions, or to other genes, located upstream or downstream of the gene contained by the isolated nucleic acid molecule when found in the chromosome. In yet another embodiment, the isolated nucleic acid lacks one or more introns. Isolated nucleic acid molecules include sequences inserted into plasmids, cosmids, artificial chromosomes, and the like. Thus, in a specific embodiment, a recombinant nucleic acid is an isolated nucleic acid. An isolated protein may be associated with other proteins or nucleic acids, or both, with which it associates in the cell, or with cellular membranes if it is a membrane-associated protein. An isolated organelle, cell, or tissue is removed from the anatomical site in which it is found in an organism. An isolated material may be, but need not be, purified.

The term "purified" as used herein refers to material that has been isolated under conditions that reduce or eliminate the presence of unrelated materials, *i.e.*, contaminants, including native materials from which the material is obtained. For example, a purified protein is preferably substantially free of other proteins or nucleic acids with which it is associated in a cell; a purified

nucleic acid molecule is preferably substantially free of proteins or other unrelated nucleic acid molecules with which it can be found within a cell. As used herein, the term "substantially free" is used operationally, in the context of analytical testing of the material. Preferably, purified material substantially free of contaminants is at least 50% pure; more preferably, at least 90% pure, and more preferably still at least 99% pure. Purity can be evaluated by chromatography, gel electrophoresis, immunoassay, composition analysis, biological assay, and other methods known in the art.

Methods for purification are well-known in the art. For example, nucleic acids can be purified by precipitation, chromatography (including preparative solid phase chromatography, oligonucleotide hybridization, and triple helix chromatography), ultracentrifugation, and other means. Polypeptides and proteins can be purified by various methods including, without limitation, preparative disc-gel electrophoresis, isoelectric focusing, HPLC, reversed-phase HPLC, gel filtration, ion exchange and partition chromatography, precipitation and salting-out chromatography, extraction, and countercurrent distribution. For some purposes, it is preferable to produce the polypeptide in a recombinant system in which the protein contains an additional sequence tag that facilitates purification, such as, but not limited to, a polyhistidine sequence, or a sequence that specifically binds to an antibody, such as FLAG and GST. The polypeptide can then be purified from a crude lysate of the host cell by chromatography on an appropriate solid-phase matrix. Alternatively, antibodies produced against the protein or against peptides derived therefrom can be used as purification reagents. Cells can be purified by various techniques, including centrifugation, matrix separation (*e.g.*, nylon wool separation), panning and other immunoselection techniques, depletion (*e.g.*, complement depletion of contaminating cells), and cell sorting (*e.g.*, fluorescence activated cell sorting [FACS]). Other purification methods are possible. A purified material may contain less than about 50%, preferably less than about 75%, and most preferably less than about 90%, of the cellular components with which it was originally associated. The "substantially pure" indicates the highest degree of purity which can be achieved using conventional purification techniques known in the art. In a specific embodiment, the term "about" or "approximately" means within 20%, preferably within 10%, and more preferably within 5% of a given value or range.

A "sample" as used herein refers to a biological material which can be tested for the presence of Ulp1 protein or *Ulp1* nucleic acids. Such samples can be obtained from animal subjects,

such as humans and non-human animals, and include tissue, biopsies, blood and blood products; plural effusions; cerebrospinal fluid (CSF); ascites fluid; and cell culture.

The term "about" or "approximately" means within an acceptable error range for a given measurement. In one aspect, the error range can be within 25% of a value, preferably within 10%, and more preferably within 5%. Alternatively, particularly in biological systems, an acceptable error is one order of magnitude, preferably 2-fold.

The use of italics indicates a nucleic acid molecule (*e.g.*, *Ulp1* cDNA, gene, etc.); normal text indicates the polypeptide or protein.

Cloning and Expression of mutant *Ulp1*

The present invention contemplates generation of a gene encoding wild-type or a mutant Ulp1, including a full length and any antigenic fragments thereof from any source, preferably human. It further contemplates detection of mutant Ulp1 protein for evaluation, diagnosis, or therapy.

In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, *e.g.*, Sambrook, Fritsch & Maniatis, *Molecular Cloning: A Laboratory Manual,* Second Edition (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (herein "Sambrook *et al.*, 1989"); *DNA Cloning: A Practical Approach,* Volumes I and II (D.N. Glover ed. 1985); *Oligonucleotide Synthesis* (M.J. Gait ed. 1984); *Nucleic Acid Hybridization* [B.D. Hames & S.J. Higgins eds. (1985)]; *Transcription And Translation* [B.D. Hames & S.J. Higgins, eds. (1984)]; *Animal Cell Culture* [R.I. Freshney, ed. (1986)]; *Immobilized Cells And Enzymes* [IRL Press, (1986)]; B.Perbal, *A Practical Guide To Molecular Cloning* (1984); F.M. Ausubel *et al.* (eds.), *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc. (1994).

Molecular Biology - Definitions

"Amplification" of DNA as used herein denotes the use of polymerase chain reaction (PCR) to increase the concentration of a particular DNA sequence within a mixture of DNA sequences. For a description of PCR see Saiki *et al.*, Science, 239:487, 1988.

The polynucleotides encoding Ulp1 herein may be flanked by natural regulatory (expression control) sequences, or may be associated with heterologous sequences, including promoters, internal ribosome entry sites (IRES) and other ribosome binding site sequences, enhancers, response elements, suppressors, signal sequences, polyadenylation sequences, introns, 5'- and 3'- non-coding regions, and the like. The nucleic acids may also be modified by many means known in the art. Non-limiting examples of such modifications include methylation, "caps", substitution of one or more of the naturally occurring nucleotides with an analog, and internucleotide modifications such as, for example, those with uncharged linkages (*e.g.*, methyl phosphonates, phosphotriesters, phosphoroamidates, carbamates, etc.) and with charged linkages (*e.g.*, phosphorothioates, phosphorodithioates, etc.). Polynucleotides may contain one or more additional covalently linked moieties, such as, for example, proteins (*e.g.*, nucleases, toxins, antibodies, signal peptides, poly-L-lysine, etc.), intercalators (*e.g.*, acridine, psoralen, etc.), chelators (*e.g.*, metals, radioactive metals, iron, oxidative metals, etc.), and alkylators. The polynucleotides may be derivatized by formation of a methyl or ethyl phosphotriester or an alkyl phosphoramidate linkage. Furthermore, the polynucleotides herein may also be modified with a label capable of providing a detectable signal, either directly or indirectly. Exemplary labels include radioisotopes, fluorescent molecules, biotin, and the like.

The term "host cell" means any cell of any organism that is selected, modified, transformed, grown, or used or manipulated in any way, for the production of a substance by the cell, for example the expression by the cell of a gene, a DNA or RNA sequence, a protein or an enzyme. Host cells can further be used for screening or other assays, as described *infra*.

Proteins and enzymes are made in the host cell using instructions in DNA and RNA, according to the genetic code. Generally, a DNA sequence having instructions for a particular protein or enzyme is "transcribed" into a corresponding sequence of RNA. The RNA sequence in turn is "translated" into the sequence of amino acids which form the protein or enzyme. An "amino acid sequence" is any chain of two or more amino acids. Each amino acid is represented in DNA or RNA by one or more triplets of nucleotides. Each triplet forms a codon, corresponding to an amino acid. For example, the amino acid lysine (Lys) can be coded by the nucleotide triplet or codon AAA or by the codon AAG. (The genetic code has some redundancy, also called degeneracy,

meaning that most amino acids have more than one corresponding codon.) Because the nucleotides in DNA and RNA sequences are read in groups of three for protein production, it is important to begin reading the sequence at the correct amino acid, so that the correct triplets are read. The way that a nucleotide sequence is grouped into codons is called the "reading frame."

A "coding sequence" or a sequence "encoding" an expression product, such as a RNA, polypeptide, protein, or enzyme, is a nucleotide sequence that, when expressed, results in the production of that RNA, polypeptide, protein, or enzyme, *i.e.,* the nucleotide sequence encodes an amino acid sequence for that polypeptide, protein or enzyme. A coding sequence for a protein may include a start codon (usually ATG) and a stop codon.

A "promoter sequence" is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. For purposes of defining the present invention, the promoter sequence is bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence will be found a transcription initiation site (conveniently defined for example, by mapping with nuclease S1), as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

A coding sequence is "under the control" or "operatively associated with" of transcriptional and translational control sequences in a cell when RNA polymerase transcribes the coding sequence into mRNA, which is then trans-RNA spliced (if it contains introns) and translated into the protein encoded by the coding sequence.

The terms "express" and "expression" mean allowing or causing the information in a gene or DNA sequence to become manifest, for example producing a protein by activating the cellular functions involved in transcription and translation of a corresponding gene or DNA sequence. A DNA sequence is expressed in or by a cell to form an "expression product" such as a protein. The expression product itself, *e.g.* the resulting protein, may also be said to be "expressed" by the cell. An expression product can be characterized as intracellular, extracellular or secreted. The term "intracellular" means something that is inside a cell. The term "extracellular" means

something that is outside a cell. A substance is "secreted" by a cell if it appears in significant measure outside the cell, from somewhere on or inside the cell.

The term "transfection" means the introduction of a foreign nucleic acid into a cell. The term "transformation" means the introduction of a "foreign" (*i.e.* extrinsic or extracellular) gene, DNA or RNA sequence to a host cell, so that the host cell will express the introduced gene or sequence to produce a desired substance, typically a protein or enzyme coded by the introduced gene or sequence. The introduced gene or sequence may also be called a "cloned" or "foreign" gene or sequence, may include regulatory or control sequences, such as start, stop, promoter, signal, secretion, or other sequences used by a cell's genetic machinery. The gene or sequence may include nonfunctional sequences or sequences with no known function. A host cell that receives and expresses introduced DNA or RNA has been "transformed" and is a "transformant" or a "clone." The DNA or RNA introduced to a host cell can come from any source, including cells of the same genus or species as the host cell, or cells of a different genus or species.

The terms "vector", "cloning vector" and "expression vector" mean the vehicle by which a DNA or RNA sequence (*e.g.* a foreign gene) can be introduced into a host cell, so as to transform the host and promote expression (*e.g.* transcription and translation) of the introduced sequence. Vectors include plasmids, phages, viruses, etc.; they are discussed in greater detail below.

Vectors typically comprise the DNA of a transmissible agent, into which foreign DNA is inserted. A common way to insert one segment of DNA into another segment of DNA involves the use of enzymes called restriction enzymes that cleave DNA at specific sites (specific groups of nucleotides) called restriction sites. A "cassette" refers to a DNA coding sequence or segment of DNA that codes for an expression product that can be inserted into a vector at defined restriction sites. The cassette restriction sites are designed to ensure insertion of the cassette in the proper reading frame. Generally, foreign DNA is inserted at one or more restriction sites of the vector DNA, and then is carried by the vector into a host cell along with the transmissible vector DNA. A segment or sequence of DNA having inserted or added DNA, such as an expression vector, can also be called a "DNA construct." A common type of vector is a "plasmid", which generally is a self-contained molecule of double-stranded DNA, usually of bacterial origin, that can readily accept additional (foreign) DNA and which can readily introduced into a suitable host cell. A

plasmid vector often contains coding DNA and promoter DNA and has one or more restriction sites suitable for inserting foreign DNA. Coding DNA is a DNA sequence that encodes a particular amino acid sequence for a particular protein or enzyme. Promoter DNA is a DNA sequence which initiates, regulates, or otherwise mediates or controls the expression of the coding DNA. Promoter DNA and coding DNA may be from the same gene or from different genes, and may be from the same or different organisms. A large number of vectors, including plasmid and fungal vectors, have been described for replication and/or expression in a variety of eukaryotic and prokaryotic hosts. Non-limiting examples include pKK plasmids (Clonetech), pUC plasmids, pET plasmids (Novagen, Inc., Madison, WI), pRSET or pREP plasmids (Invitrogen, San Diego, CA), or pMAL plasmids (New England Biolabs, Beverly, MA), and many appropriate host cells, using methods disclosed or cited herein or otherwise known to those skilled in the relevant art. Recombinant cloning vectors will often include one or more replication systems for cloning or expression, one or more markers for selection in the host, *e.g.* antibiotic resistance, and one or more expression cassettes.

The term "expression system" means a host cell and compatible vector under suitable conditions, *e.g.* for the expression of a protein coded for by foreign DNA carried by the vector and introduced to the host cell. Common expression systems include *E. coli* host cells and plasmid vectors, insect host cells and *Baculovirus* vectors, and mammalian host cells and vectors.

The term "heterologous" refers to a combination of elements not naturally occurring. For example, heterologous DNA refers to DNA not naturally located in the cell, or in a chromosomal site of the cell. Preferably, the heterologous DNA includes a gene foreign to the cell. A heterologous expression regulatory element is a such an element operatively associated with a different gene than the one it is operatively associated with in nature. In the context of the present invention, a *Ulp1* gene is heterologous to the vector DNA in which it is inserted for cloning or expression, and it is heterologous to a host cell containing such a vector, in which it is expressed, *e.g.*, a yeast cell.

The terms "mutant" and "mutation" mean any detectable change in genetic material, *e.g.* DNA, or any process, mechanism, or result of such a change. This includes gene mutations, in which the structure (*e.g.* DNA sequence) of a gene is altered, any gene or DNA arising from any mutation process, and any expression product (*e.g.* protein or enzyme) expressed by a modified gene

or DNA sequence. The term "variant" may also be used to indicate a modified or altered gene, DNA sequence, enzyme, cell, etc., *i.e.*, any kind of mutant.

"Sequence-conservative variants" of a polynucleotide sequence are those in which a change of one or more nucleotides in a given codon position results in no alteration in the amino acid encoded at that position.

"Function-conservative variants" are those in which a given amino acid residue in a protein or enzyme has been changed without altering the overall conformation and function of the polypeptide, including, but not limited to, replacement of an amino acid with one having similar properties (such as, for example, polarity, hydrogen bonding potential, acidic, basic, hydrophobic, aromatic, and the like). Amino acids with similar properties are well known in the art. For example, arginine, histidine and lysine are hydrophilic-basic amino acids and may be interchangeable. Similarly, isoleucine, a hydrophobic amino acid, may be replaced with leucine, methionine or valine. Such changes are expected to have little or no effect on the apparent molecular weight or isoelectric point of the protein or polypeptide. Amino acids other than those indicated as conserved may differ in a protein or enzyme so that the percent protein or amino acid sequence similarity between any two proteins of similar function may vary and may be, for example, from 70% to 99% as determined according to an alignment scheme such as by the Cluster Method, wherein similarity is based on the MEGALIGN algorithm. A "function-conservative variant" also includes a polypeptide or enzyme which has at least 60 % amino acid identity as determined by BLAST or FASTA algorithms, preferably at least 75%, most preferably at least 85%, and even more preferably at least 90%, and which has the same or substantially similar properties or functions as the native or parent protein or enzyme to which it is compared.

As used herein, the term "homologous" in all its grammatical forms and spelling variations refers to the relationship between proteins that possess a "common evolutionary origin," including proteins from superfamilies (*e.g.*, the immunoglobulin superfamily) and homologous proteins from different species (*e.g.*, myosin light chain, etc.) (Reeck *et al.*, 1987, Cell 50:667). Such proteins (and their encoding genes) have sequence homology, as reflected by their sequence similarity, whether in terms of percent similarity or the presence of specific residues or motifs at conserved positions.

Accordingly, the term "sequence similarity" in all its grammatical forms refers to the degree of identity or correspondence between nucleic acid or amino acid sequences of proteins that may or may not share a common evolutionary origin (*see* Reeck *et al.*, *supra*). However, in common usage and in the instant application, the term "homologous," when modified with an adverb such as "highly," may refer to sequence similarity and may or may not relate to a common evolutionary origin.

In a specific embodiment, two DNA sequences are "substantially homologous" or "substantially similar" when at least about 80%, and most preferably at least about 90 or 95%) of the nucleotides match over the defined length of the DNA sequences, as determined by sequence comparison algorithms, such as BLAST, FASTA, DNA Strider, etc. An example of such a sequence is an allelic or species variant of the specific *Ulp1* genes of the invention. Sequences that are substantially homologous can be identified by comparing the sequences using standard software available in sequence data banks, or in a Southern hybridization experiment under, for example, stringent conditions as defined for that particular system.

Similarly, in a particular embodiment, two amino acid sequences are "substantially homologous" or "substantially similar" when greater than 80% of the amino acids are identical, or greater than about 90% are similar (functionally identical). Preferably, the similar or homologous sequences are identified by alignment using, for example, the GCG (Genetics Computer Group, Program Manual for the GCG Package, *Version 7*, Madison, Wisconsin) pileup program, or any of the programs described above (BLAST, FASTA, etc)

A nucleic acid molecule is "hybridizable" to another nucleic acid molecule, such as a cDNA, genomic DNA, or RNA, when a single stranded form of the nucleic acid molecule can anneal to the other nucleic acid molecule under the appropriate conditions of temperature and solution ionic strength (*see* Sambrook *et al.*, *supra*). The conditions of temperature and ionic strength determine the "stringency" of the hybridization. For preliminary screening for homologous nucleic acids, low stringency hybridization conditions, corresponding to a T_m (melting temperature) of 55°C, can be used, *e.g.*, 5x SSC, 0.1% SDS, 0.25% milk, and no formamide; or 30% formamide, 5x SSC, 0.5% SDS). Moderate stringency hybridization conditions correspond to a higher T_m, *e.g.*, 40% formamide, with 5x or 6x SCC. High stringency hybridization conditions correspond to the

highest T_m, *e.g.*, 50% formamide, 5x or 6x SCC. SCC is a 0.15M NaCl, 0.015M Na-citrate. Hybridization requires that the two nucleic acids contain complementary sequences, although depending on the stringency of the hybridization, mismatches between bases are possible. The appropriate stringency for hybridizing nucleic acids depends on the length of the nucleic acids and the degree of complementation, variables well known in the art. The greater the degree of similarity or homology between two nucleotide sequences, the greater the value of T_m for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA. For hybrids of greater than 100 nucleotides in length, equations for calculating T_m have been derived (*see* Sambrook *et al.*, *supra*, 9.50-9.51). For hybridization with shorter nucleic acids, *i.e.*, oligonucleotides, the position of mismatches becomes more important, and the length of the oligonucleotide determines its specificity (*see* Sambrook *et al.*, *supra*, 11.7-11.8). A minimum length for a hybridizable nucleic acid is at least about 10 nucleotides; preferably at least about 15 nucleotides; and more preferably the length is at least about 20 nucleotides.

In a specific embodiment, the term "standard hybridization conditions" refers to a T_m of 55°C, and utilizes conditions as set forth above. In a preferred embodiment, the T_m is 60°C; in a more preferred embodiment, the T_m is 65°C. In a specific embodiment, "high stringency" refers to hybridization and/or washing conditions at 68°C in 0.2XSSC, at 42°C in 50% formamide, 4XSSC, or under conditions that afford levels of hybridization equivalent to those observed under either of these two conditions.

As used herein, the term "oligonucleotide" refers to a nucleic acid, generally of at least 10, preferably at least 15, and more preferably at least 20 nucleotides, preferably no more than 100 nucleotides, that is hybridizable to a genomic DNA molecule, a cDNA molecule, or an mRNA molecule encoding a gene, mRNA, cDNA, or other nucleic acid of interest. Oligonucleotides can be labeled, *e.g.*, with ^{32}P-nucleotides or nucleotides to which a label, such as biotin, has been covalently conjugated. In one embodiment, a labeled oligonucleotide can be used as a probe to detect the presence of a nucleic acid. In another embodiment, oligonucleotides (one or both of which may be labeled) can be used as PCR primers, either for cloning full length or a fragment of Ulp1, or to detect the presence of nucleic acids encoding Ulp1. In a further embodiment, an

oligonucleotide of the invention can form a triple helix with a Ulp1 DNA molecule. Generally, oligonucleotides are prepared synthetically, preferably on a nucleic acid synthesizer. Accordingly, oligonucleotides can be prepared with non-naturally occurring phosphoester analog bonds, such as thioester bonds, etc.

Ulp1 Nucleic Acids

A gene encoding mutant Ulp1, whether genomic DNA or cDNA, can be isolated from any source, particularly from a human cDNA or genomic library. Methods for obtaining *Ulp1* gene are well known in the art, as described above (*see, e.g.*, Sambrook *et al.*, 1989, *supra*). The DNA may be obtained by standard procedures known in the art from cloned DNA (*e.g.*, a DNA "library"), and preferably is obtained from a cDNA library prepared from tissues with high level expression of the protein, by chemical synthesis, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from the desired cell (See, for example, Sambrook *et al.*, 1989, *supra*; Glover, D.M. (ed.), 1985, DNA Cloning: A Practical Approach, MRL Press, Ltd., Oxford, U.K. Vol. I, II). Clones derived from genomic DNA may contain regulatory and intron DNA regions in addition to coding regions; clones derived from cDNA will not contain intron sequences. Whatever the source, the gene should be molecularly cloned into a suitable vector for propagation of the gene. Identification of the specific DNA fragment containing the desired *Ulp1* gene may be accomplished in a number of ways. For example, a portion of an *Ulp1* gene exemplified *infra* can be purified and labeled to prepare a labeled probe, and the generated DNA may be screened by nucleic acid hybridization to the labeled probe (Benton and Davis, Science 196:180, 1977; Grunstein and Hogness, Proc. Natl. Acad. Sci. U.S.A. 72:3961, 1975). Those DNA fragments with substantial homology to the probe, such as an allelic variant from another individual, will hybridize.

Further selection can be carried out on the basis of the properties of the gene, *e.g.*, if the gene encodes a protein product having the isoelectric, electrophoretic, amino acid composition, partial or complete amino acid sequence, antibody binding activity, or ligand binding profile of *Ulp1* protein as disclosed herein. Thus, the presence of the gene may be detected by assays based on the physical, chemical, immunological, or functional properties of its expressed product.

Other DNA sequences which encode substantially the same amino acid sequence as an *Ulp1* gene may be used in the practice of the present invention. These include but are not limited to allelic variants, species variants, sequence conservative variants, and functional variants.

Amino acid substitutions may also be introduced to substitute an amino acid with a particularly preferable property. For example, a Cys may be introduced a potential site for disulfide bridges with another Cys.

The genes encoding *Ulp1* derivatives and analogs of the invention can be produced by various methods known in the art. The manipulations which result in their production can occur at the gene or protein level. For example, the cloned *Ulp1* gene sequence can be modified by any of numerous strategies known in the art (Sambrook *et al.*, 1989, *supra*). The sequence can be cleaved at appropriate sites with restriction endonuclease(s), followed by further enzymatic modification if desired, isolated, and ligated *in vitro*. In the production of the gene encoding a derivative or analog of *Ulp1*, care should be taken to ensure that the modified gene remains within the same translational reading frame as the *Ulp1* gene, uninterrupted by translational stop signals, in the gene region where the desired activity is encoded.

Additionally, the *Ulp1* encoding nucleic acid sequence can be mutated *in vitro* or *in vivo*, to create and/or destroy translation, initiation, and/or termination sequences, or to create variations in coding regions and/or form new restriction endonuclease sites or destroy preexisting ones, to facilitate further *in vitro* modification. Such modifications can be made to introduce restriction sites and facilitate cloning the *Ulp1* gene into an expression vector. Any technique for mutagenesis known in the art can be used, including but not limited to, *in vitro* site-directed mutagenesis (Hutchinson, C., *et al.*, J. Biol. Chem. 253:6551, 1978; Zoller and Smith, DNA 3:479-488, 1984; Oliphant *et al.*, Gene 44:177, 1986; Hutchinson *et al.*, Proc. Natl. Acad. Sci. U.S.A. 83:710, 1986), use of TAB" linkers (Pharmacia), etc. PCR techniques are preferred for site directed mutagenesis (see Higuchi, 1989, "Using PCR to Engineer DNA", in *PCR Technology: Principles and Applications for DNA Amplification*, H. Erlich, ed., Stockton Press, Chapter 6, pp. 61-70).

The identified and isolated gene can then be inserted into an appropriate cloning vector. A large number of vector-host systems known in the art may be used. Possible vectors include, but are not limited to, plasmids or modified viruses, but the vector system must be

compatible with the host cell used. Examples of vectors include, but are not limited to, *E. coli*, bacteriophages such as lambda derivatives, or plasmids such as pBR322 derivatives or pUC plasmid derivatives, *e.g.*, pGEX vectors, pmal-c, pFLAG, etc. The insertion into a cloning vector can, for example, be accomplished by ligating the DNA fragment into a cloning vector which has complementary cohesive termini. However, if the complementary restriction sites used to fragment the DNA are not present in the cloning vector, the ends of the DNA molecules may be enzymatically modified. Alternatively, any site desired may be produced by ligating nucleotide sequences (linkers) onto the DNA termini; these ligated linkers may comprise specific chemically synthesized oligonucleotides encoding restriction endonuclease recognition sequences.

Recombinant molecules can be introduced into host cells via transformation, transfection, infection, electroporation, etc., so that many copies of the gene sequence are generated. Preferably, the cloned gene is contained on a shuttle vector plasmid, which provides for expansion in a cloning cell, *e.g.*, *E. coli*, and facile purification for subsequent insertion into an appropriate expression cell line, if such is desired. For example, a shuttle vector, which is a vector that can replicate in more than one type of organism, can be prepared for replication in both *E. coli* and *Saccharomyces cerevisiae* by linking sequences from an *E. coli* plasmid with sequences form the yeast 2μ plasmid.

Expression of Ulp1 Polypeptides

The nucleotide sequence coding for Ulp1, or antigenic fragment, derivative or analog thereof, or a functionally active derivative, including a chimeric protein, thereof, can be inserted into an appropriate expression vector, *i.e.*, a vector which contains the necessary elements for the transcription and translation of the inserted protein-coding sequence. Thus, a nucleic acid encoding Ulp1 of the invention can be operationally associated with a promoter in an expression vector of the invention. Both cDNA and genomic sequences can be cloned and expressed under control of such regulatory sequences. Such vectors can be used to express functional or functionally inactivated Ulp1 polypeptides.

The necessary transcriptional and translational signals can be provided on a recombinant expression vector, or they may be supplied by the native gene encoding Ulp1 and/or its flanking regions.

Potential host-vector systems include but are not limited to mammalian cell systems transfected with expression plasmids or infected with virus (*e.g.*, vaccinia virus, adenovirus, adeno-associated virus, herpes virus, etc.); insect cell systems infected with virus (*e.g.*, baculovirus); microorganisms such as yeast containing yeast vectors; or bacteria transformed with bacteriophage, DNA, plasmid DNA, or cosmid DNA. The expression elements of vectors vary in their strengths and specificities. Depending on the host-vector system utilized, any one of a number of suitable transcription and translation elements may be used.

Expression of Ulp1 protein may be controlled by any promoter/enhancer element known in the art, but these regulatory elements must be functional in the host selected for expression. Promoters which may be used to control Ulp1 gene expression include, but are not limited to, cytomegalovirus (CMV) promoter, the SV40 early promoter region (Benoist and Chambon, 1981, Nature 290:304-310), the promoter contained in the 3' long terminal repeat of Rous sarcoma virus (Yamamoto, *et al.*, Cell 22:787-797, 1980), the herpes thymidine kinase promoter (Wagner *et al.*, Proc. Natl. Acad. Sci. U.S.A. 78:1441-1445, 1981), the regulatory sequences of the metallothionein gene (Brinster *et al.*, Nature 296:39-42, 1982); prokaryotic expression vectors such as the β-lactamase promoter (Villa-Komaroff, *et al.*, Proc. Natl. Acad. Sci. U.S.A. 75:3727-3731, 1978), or the *tac* promoter (DeBoer, *et al.*, Proc. Natl. Acad. Sci. U.S.A. 80:21-25, 1983); see also "Useful proteins from recombinant bacteria" in Scientific American, 242:74-94, 1980; promoter elements from yeast or other fungi such as the Gal 4 promoter, the ADC (alcohol dehydrogenase) promoter, PGK (phosphoglycerol kinase) promoter, alkaline phosphatase promoter; and transcriptional control regions that exhibit tissue specificity, particularly endothelial cell-specific promoters.

Soluble forms of the protein can be obtained by collecting culture fluid, or solubilizing inclusion bodies, *e.g.*, by treatment with detergent, and if desired sonication or other mechanical processes, as described above. The solubilized or soluble protein can be isolated using various techniques, such as polyacrylamide gel electrophoresis (PAGE), isoelectric focusing, 2-dimensional gel electrophoresis, chromatography (*e.g.*, ion exchange, affinity, immunoaffinity,

and sizing column chromatography), centrifugation, differential solubility, immunoprecipitation, or by any other standard technique for the purification of proteins.

Vectors

A wide variety of host/expression vector combinations may be employed in expressing the DNA sequences of this invention. Useful expression vectors, for example, may consist of segments of chromosomal, non-chromosomal and synthetic DNA sequences. Suitable vectors include derivatives of SV40 and known bacterial plasmids, *e.g.*, *E. coli* plasmids col El, pCR1, pBR322, pMal-C2, pET, pGEX (Smith *et al.*, Gene 67:31-40, 1988), pMB9 and their derivatives, plasmids such as RP4; phage DNAS, *e.g.*, the numerous derivatives of phage l, *e.g.*, NM989, and other phage DNA, *e.g.*, M13 and filamentous single stranded phage DNA; yeast plasmids such as the 2μ plasmid or derivatives thereof; vectors useful in eukaryotic cells, such as vectors useful in insect or mammalian cells; vectors derived from combinations of plasmids and phage DNAs, such as plasmids that have been modified to employ phage DNA or other expression control sequences; and the like.

Preferred vectors for introducing Ulp1 coding sequences into mammalian or insect cells are viral vectors, such as lentiviruses, retroviruses, herpes viruses, adenoviruses, adeno-associated viruses, vaccinia virus, baculovirus, and other recombinant viruses with desirable cellular tropism. Thus, a gene encoding a functional or mutant Ulp1 protein or polypeptide domain fragment thereof can be introduced using a viral vector or through direct introduction of DNA.

Viral vectors commonly used for targeting procedures are DNA-based vectors and retroviral vectors. Methods for constructing and using viral vectors are known in the art (*see, e.g.*, Miller and Rosman, BioTechniques, 7:980-990, 1992). Preferably, the viral vectors are replication defective, that is, they are unable to replicate autonomously in the target cell. In general, the genome of the replication defective viral vectors which are used within the scope of the present invention lack at least one region which is necessary for the replication of the virus in the infected cell. These regions can either be eliminated (in whole or in part), be rendered non-functional by any technique known to a person skilled in the art. These techniques include the total removal, substitution (by other sequences, in particular by the inserted nucleic acid), partial deletion or addition of one or more

bases to an essential (for replication) region. Such techniques may be performed *in vitro* (on the isolated DNA) or *in situ*, using the techniques of genetic manipulation or by treatment with mutagenic agents. Preferably, the replication defective virus retains the sequences of its genome which are necessary for encapsidating the viral particles.

DNA viral vectors include an attenuated or defective DNA virus, such as but not limited to herpes simplex virus (HSV), papillomavirus, Epstein Barr virus (EBV), adenovirus, adeno-associated virus (AAV), and the like. Defective viruses, which entirely or almost entirely lack viral genes, are preferred. Defective virus is not infective after introduction into a cell. Use of defective viral vectors allows for administration to cells in a specific, localized area, without concern that the vector can infect other cells. Thus, a specific tissue can be specifically targeted. Examples of particular vectors include, but are not limited to, a defective herpes virus 1 (HSV1) vector (Kaplitt *et al.*, Molec. Cell. Neurosci. 2:320-330, 1991), defective herpes virus vector lacking a glyco-protein L gene (Patent Publication RD 371005 A), or other defective herpes virus vectors (International Patent Publication No. WO 94/21807, published September 29, 1994; International Patent Publication No. WO 92/05263, published April 2, 1994); an attenuated adenovirus vector, such as the vector described by Stratford-Perricaudet *et al.* (J. Clin. Invest. 90:626-630, 1992; see also La Salle *et al.*, Science 259:988-990, 1993); and a defective adeno-associated virus vector (Samulski *et al.*, J. Virol. 61:3096-3101, 1987; Samulski *et al.*, J. Virol. 63:3822-3828, 1989; Lebkowski *et al.*, Mol. Cell. Biol. 8:3988-3996, 1988).

Various companies produce viral vectors commercially, including but by no means limited to Avigen, Inc. (Alameda, CA; AAV vectors), Cell Genesys (Foster City, CA; retroviral, adenoviral, AAV vectors, and lentiviral vectors), Clontech (retroviral and baculoviral vectors), Genovo, Inc. (Sharon Hill, PA; adenoviral and AAV vectors), Genvec (adenoviral vectors), IntroGene (Leiden, Netherlands; adenoviral vectors), Molecular Medicine (retroviral, adenoviral, AAV, and herpes viral vectors), Norgen (adenoviral vectors), Oxford BioMedica (Oxford, United Kingdom; lentiviral vectors), and Transgene (Strasbourg, France; adenoviral, vaccinia, retroviral, and lentiviral vectors).

In another embodiment, the vector can be introduced by lipofection, as naked DNA, or with other transfection facilitating agents (peptides, polymers, etc.). Synthetic cationic lipids can

be used to prepare liposomes for *in vivo* transfection of a gene encoding a marker (Felgner, et. al., Proc. Natl. Acad. Sci. U.S.A. 84:7413-7417, 1987; Felgner and Ringold, Science 337:387-388, 1989; see Mackey, *et al.*, Proc. Natl. Acad. Sci. U.S.A. 85:8027-8031, 1988; Ulmer, *et al.*, Science 259:1745-1748, 1993). Useful lipid compounds and compositions for transfer of nucleic acids are described in International Patent Publications WO95/18863 and WO96/17823, and in U.S. Patent No. 5,459,127. Lipids may be chemically coupled to other molecules for the purpose of targeting (see Mackey, *et al.*, *supra*). Targeted peptides, *e.g.*, hormones or neurotransmitters, and proteins such as antibodies, or non-peptide molecules could be coupled to liposomes chemically.

Other molecules are also useful for facilitating transfection of a nucleic acid, such as a cationic oligopeptide (*e.g.*, International Patent Publication WO95/21931), peptides derived from DNA binding proteins (*e.g.*, International Patent Publication WO96/25508), or a cationic polymer (*e.g.*, International Patent Publication WO95/21931).

Alternatively, non-viral DNA vectors can be introduced into the desired host cells by methods known in the art, *e.g.*, electroporation, microinjection, cell fusion, DEAE dextran, calcium phosphate precipitation, use of a gene gun (ballistic transfection; see, *e.g.*, U.S. Pat. No. 5,204,253, U.S. Pat. No. 5,853,663, U.S. Pat. No. 5,885,795, and U.S. Pat. No. 5,702,384 and see Sanford, TIB-TECH, 6:299-302, 1988; Fynan *et al.*, Proc. Natl. Acad. Sci. U.S.A., 90:11478-11482, 1993; and Yang *et al.*, Proc. Natl. Acad. Sci. U.S.A., 87:1568-9572, 1990), or use of a DNA vector transporter (see, *e.g.*, Wu, *et al.*, J. Biol. Chem. 267:963-967, 1992; Wu and Wu, J. Biol. Chem. 263:14621-14624, 1988; Hartmut, *et al.*, Canadian Patent Application No. 2,012,311, filed March 15, 1990; Williams, *et al.*, Proc. Natl. Acad. Sci. USA 88:2726-2730, 1991). Receptor-mediated DNA delivery approaches can also be used (Curiel, *et al.*, Hum. Gene Ther. 3:147-154, 1992; Wu and Wu, J. Biol. Chem. 262:4429-4432, 1987).

Protease-Substrate Complex Formation

The present invention describes a composition comprising a polypeptide of a SUMO protease catalytic domain in a trapped proteolytic deacylation intermediate complex with its substrate. The composition may then be treated with methods known in the art to produce a crystalline structure of the complex. The present invention contemplates characterization of the structure of any protease that produces deacylation intermediates upon interaction with a substrate, particularly any cysteine protease, *e.g.*, by trapping the protease-substrate covalent intermediate. Specifically, the present invention contemplates a composition where the SUMO protease is Ulp1. Ulp1 may be derived from various sources, including, but not limited to, mammal, yeast, insects, and plants. In one preferred embodiment, the Ulp1 is derived from mammal, specifically human. The composition may include any substrate of Ulp1, such as SUMO and Smt3. In a specific embodiment, the substrate is Smt3.

The complexed molecules in the composition are preferably obtained in a crystalline structure. The complexed molecules may be formed into a crystalline structure using techniques that are well known in the art, where the crystallizing conditions are modified and optimized for crystal formation of specific proteases and their substrate. Alternatively, other spectroscopic techniques for structure evaluation, such as nuclear magnetic resonance, Raman spectroscopy, circular dichroism, and neutron diffraction can be used to study trapped protease-substrate complexes.

The coordinates of atoms and structure of the protease and substrate in the crystal structure may be defined using any method and computational method needed. For example, if the protease and substrate complex are in a liquid form, then nuclear magnetic resonance can be used to determine the coordinates and structure of the complex. For complexes that are in a crystalline structure x-ray crystallography may be used to elucidate the structure.

The invention also describes a method of preparatively trapping and purifying large amounts of the proteolytic deacylation intermediate using a reducing agent. The method of trapping the proteolytic deacylation intermediate involves (a) combining the protease and substrate in a molar ratio, (b) adding a reducing agent that is capable of trapping the substrate in an intermediate state with the protease, and (c) then adjusting the pH of the solution to about 7.0. The invention allows for performing steps (a) and (b) in any order or simultaneously. In a

specific embodiment the protease is a SUMO protease, and more specifically the SUMO protease is the catalytic domain of Ulp1. Ulp1 may be derived from various sources, including, but not limited to, mammal, yeast, insects, and plants. In one preferred embodiment, the Ulp1 is derived from mammal, specifically human. The complex includes any substrate of Ulp1, such as SUMO and Smt3. In a specific embodiment, the substrate is Smt3.

The pH of the solution may be adjusted to about 7.0 by various methods known in the art, such as adding the required amount of an acid or base to the solution to adjust the pH to about 7.0. Other methods that may be used to adjust the pH of a solution, as appropriate for the system, include, but not limited to, chromatography and dialysis. In one embodiment of the invention, the pH of the solution is adjusted by dialysis. The protease-substrate complex may then be isolated from the solution after the pH is adjusted by dialysis. The protease-substrate complex may be isolated from the solution using any methods known in the art, including high performance liquid chromatography, size exclusion chromatography, dialysis, fast protein liquid chromatography, and crystallization. Other method of isolating the complex may be used, depending on the protease and substrate.

The molar ratio of protease to substrate may encompass any ratio where the concentration of substrate is in excess to insure that the protease may form a complex with it. Theoretically, the substrate may be in infinite excess to the protease, however that would lead to significant waste of substrate that remains uncomplexed to the protease. In the present invention the molar ratio of the protease to substrate ranges from 1:1 to 1:5. More preferred, the Ulp1 and Smt3 are in a 1:3 molar ratio.

The protease and substrate are trapped in a deacylation complex by addition of a reducing agent in an amount that is effective to trap the proteins. An effective amount of reducing agent is defined as the amount needed to trap an isolatable amount of the complex in the solution. Any reducing agent is contemplated by the present invention. However, several reducing agents, such as lithium aluminum hydride, could reduce other portions of the protease and substrate in the complex. In one specific embodiment the reducing agent is sodium borohydride.

Screening and Chemistry

According to the present invention, the structure of a Ulp1-Smt3 complex, or other trapped protease-substrate complexes, are useful to identify drugs that are effective in treating disorders associated with Ub/Ubl pathway, including tumorogenesis and acute promyelocytic leukemia.

Rational Drug Design

The invention also defines a method of identifying novel substrates for cysteine proteases by using rational drug design methods. The invention allows for designing substrates that would interact with amino acids in the protease catalytic site as defined by computer molecular modeling methods based on the crystal structure of the protease and other substrates. A substrate may include a competitive inhibitor, analog of the native substrate, or a suicide inhibitor. The designed substrates would be designed so that they interact with amino acid residues that are defined to be (i) necessary for interaction with a substrate and (ii) protease activity.

The present invention contemplates evaluating potential compounds for covalent and non-covalent interactions between the protease and substrate. Computer modeling methods that may be used to evaluate these interactions include, but are not limited to, SYBYL and Monte Carlo computer programs. The present invention contemplates computer algorithms that evaluate bonded and non-bonded interactions between the protease and the substrate. Bonded interactions that may be evaluated include, but are not limited to, bond stretching, rotational strain, and torsional strain. Non-bonded interactions that may be evaluated include van Der Waals forces, hydrogen bonds and dipole-dipole interaction.

In particular, identification of putative Ulp binding compounds provides for development of screening assays, particularly for high throughput screening of molecules that up- or down-regulate the activity of Ulp1, *e.g.*, by permitting expression of Ulp1 in quantities greater than can be isolated from natural sources, or in indicator cells that are specially engineered to indicate the activity of Ulp1 expressed after transfection or transformation of the cells.

Any screening technique known in the art can be used to screen for Ulp1 agonists or antagonists. The present invention contemplates screens for small molecule ligands or ligand analogs and mimics, as well as screens for natural ligands that bind to and agonize or antagonize

Ulp1 expression activity *in vivo*. For example, natural products libraries can be screened using assays of the invention for molecules that agonize or antagonize Ulp1 expression or activity.

Another approach uses recombinant bacteriophage to produce large libraries. Using the "phage method" (Scott and Smith, Science 249:386-390, 1990; Cwirla, *et al.*, Proc. Natl. Acad. Sci., 87:6378-6382, 1990; Devlin *et al.*, Science, 49:404-406, 1990), very large libraries can be constructed (10^6-10^8 chemical entities). A second approach uses primarily chemical methods, of which the Geysen method (Geysen *et al.*, Molecular Immunology 23:709-715, 1986; Geysen *et al.* J. Immunologic Method 102:259-274, 1987; and the method of Fodor *et al.* (Science 251:767-773, 1991) are examples. Furka *et al.* (14th International Congress of Biochemistry, Volume #5, Abstract FR:013, 1988; Furka, Int. J. Peptide Protein Res. 37:487-493, 1991), Houghton (U.S. Patent No. 4,631,211, issued December 1986) and Rutter *et al.* (U.S. Patent No. 5,010,175, issued April 23, 1991) describe methods to produce a mixture of peptides that can be tested as agonists or antagonists.

In another aspect, synthetic libraries (Needels *et al.*, Proc. Natl. Acad. Sci. USA 90:10700-4, 1993; Ohlmeyer *et al.*, Proc. Natl. Acad. Sci. USA 90:10922-10926, 1993; Lam *et al.*, International Patent Publication No. WO 92/00252; Kocis *et al.*, International Patent Publication No. WO 9428028) and the like can be used to screen for Ulp1 ligands according to the present invention.

Test compounds are screened from large libraries of synthetic or natural compounds. Numerous means are currently used for random and directed synthesis of saccharide, peptide, and nucleic acid based compounds. Synthetic compound libraries are commercially available from Maybridge Chemical Co. (Trevillet, Cornwall, UK), Comgenex (Princeton, NJ), Brandon Associates (Merrimack, NH), and Microsource (New Milford, CT). A rare chemical library is available from Aldrich (Milwaukee, WI). Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and animal extracts are available from *e.g.* Pan Laboratories (Bothell, WA) or MycoSearch (NC), or are readily producible. Additionally, natural and synthetically produced libraries and compounds are readily modified through conventional chemical, physical, and biochemical means (Blondelle *et al.*, Tib Tech, 14:60, 1996).

Knowledge of the crystal structure of Ulp1, particularly when trapped with a substrate, can provide an initial clue as the inhibitors or antagonists of the protein. Moreover, based

on sequence and enzymatic activity similarities of Ulp1 to other cysteine proteases, particularly cell-cycle associated cysteine proteases, the crystal structure of Ulp1 permits development of classes of cysteine protease inhibitors. Identification and screening of antagonists is further facilitated by determining structural features of the protein, *e.g.*, using X-ray crystallography, neutron diffraction, nuclear magnetic resonance spectrometry, and other techniques for structure determination. These techniques provide for the rational design or identification of agonists and antagonists.

In vivo *screening methods*

Intact cells or whole animals expressing a gene encoding Ulp1 can be used in screening methods to identify candidate drugs.

In one series of embodiments, a permanent cell line is established. Alternatively, cells (including without limitation mammalian, insect, yeast, or bacterial cells) are transiently programmed to express an *Ulp1* gene by introduction of appropriate DNA or mRNA. Identification of candidate compounds can be achieved using any suitable assay, including without limitation (i) assays that measure selective binding of test compounds to Ulp1 (ii) assays that measure the ability of a test compound to modify (*i.e.*, inhibit or enhance) a measurable activity or function of Ulp1 and (iii) assays that measure the ability of a compound to modify (*i.e.*, inhibit or enhance) the transcriptional activity of sequences derived from the promoter (*i.e.*, regulatory) regions the Ulp1 gene.

High-Throughput Screen

Agents according to the invention may be identified by screening in high-throughput assays, including without limitation cell-based or cell-free assays. It will be appreciated by those skilled in the art that different types of assays can be used to detect different types of agents. Several methods of automated assays have been developed in recent years so as to permit screening of tens of thousands of compounds in a short period of time. Such high-throughput screening methods are particularly preferred. The use of high-throughput screening assays to test for agents is greatly facilitated by the availability of large amounts of purified polypeptides, as provided by the invention.

EXAMPLES

The present invention will be better understood by reference to the following Examples, which are provided as exemplary of the invention, and not by way of limitation.

Materials and Methods

Protein purification

A Ulp1 amino acid fragment, from amino acids 403-621 (Ulp1(403-621)) , was amplified by polymerase chain reaction. The PCR fragment was digested with Nhe1 and Xho1 restriction endonucleases and then ligated into a pET-28b vector (Novagen). *E. coli* BL21 (DE3)pLysS bacterial cells were transformed with Ulp1(403-621). Cells were grown at 37° C for 20-24 hours. Cell cultures then were grown at 37° C in superbroth, containing 50 mg/ml kanamycin and 25 mg/ml chloamphenicol, by fermentation. When the A_{600} of the culture reached about 2.0 the culture was cooled to 30° C, adjusted to 1.0 mM IPTG, and fermented for about 4 hours at 30° C. Remaining steps were performed at 4° C. The culture was centrifuged and the cell pellet was resuspended in 200 mL lysis buffer (50 mM Tris-HCl, 0.15 mM NaCl, 20% (w/v) sucrose; pH = 8.0) and sonicated. Insoluble material was removed by centrifugation and the soluble fraction was placed over 10 mL of Ni-NTA-agarose resin (Qiagen) and washed with buffer A (50 mM Tris-HCl, 0.5 M NaCl, 20 mM imidazole). The protein was eluted from the resin with a gradient (20-300 mM imidazole) in buffer A. Peak fractions of His-tagged Ulp1(403-621)p were pooled and dialyzed against 50 mM Tris-HCl (pH=8.0) containing 200 mM NaCl, 2 mM β-mercaptoethanol, and 100 units bovine thrombin (Sigma) for about 20-24 hours.

A Smt3 PCR fragment (Smt3(13-101)p) was digested with Nco1 and Xho1 restriction endonucleases and then ligated into a pET-28b vector. *E. coli* BL21 (DE3)pLysS bacterial cells were transformed with Smt3(13-101)p. Expression and purification of the protein was performed following the same protocol as described for Ulp1(403-621)p, with the exception that the protein was dialyzed against 50 mM Tris-HCl (pH=8.0) containing 200 mM NaCl.

Sumo (??-101)p was prepared in a similar manner to that described for Smt3(13-101)p.

Selenomethionyl-containing Ulp1(403-621) was expressed in DL41 cells (methionine auxotrophe) that contained the bacteriophage T7 polymerase (Hendrickson *et al.*, EMBO J., 1990, 9:1665-1672).

The Smt3-GFPuv fusion protein was constructed by ligating SMT3 into the pGFPuv vector (Clontech). Smt3-GFPuv was overexpressed in JM109 *E. coli* strain. Purification was attained by standard chromatographic techniques and included passes over size exclusion, anion-exchange resins, and finally, application to a MonoQ column (Pharmacia). The Smt3-GFP moiety was followed during purification by analyzing fluorescence over a standard UV transilluminator.

His6-ubiquitin-Smt3-HA was expressed from pQE30 (Li and Hochstrasser) in JM109 *E. coli*. The histidine-tagged fusion protein (His6-ubiquitin-Smt3-HA) was placed over Ni-agarose resin, washed and eluted as previously described. The pooled fractions were then placed onto a gel filtration column (Superdex 75 prep-grade) for size exclusion. These fractions were pooled and placed onto a MonoQ column (Pharmacia) and eluted with a 0.02-1M gradient of NaCl. Smt3-GFP, Sumo-His, and His6-ubiquitin-Smt3-HA were concentrated to ~10 mg/ml and frozen at -80° C.

Proteolysis assays

Proteolysis assays were prepared by incubating a solution at about 37° C that contained about 1.0 mg/ml substrate and about 10^{-3}-10^{-6} mg/ml Ulp1(403-621)p in buffer containing 20mM Tris-HCl (pH = 8.0) and 150mM NaCl. The Smt3-GFP reaction was subsequently analyzed by native-gel electrophoresis. The reactions containing Sumo-His and His6-ubiquitin-Smt3-HA were analyzed by SDS-PAGE.

Synthesis of covalent adduct between Ulp1(403-621)p and Smt3(13-101)p

Ulp1(403-621)p and Smt3(13-101)p were placed in a beaker at a 1:3 molar ratio with stirring. Then 5 aliquots of sodium borohydride were stirred into the beaker over 30 minutes to a final concentration of 50 mM. The mixture was dialyzed against 20 mM Tris-HCl (pH=8.0) containing 20 mM NaCl at 4° C. The dialyzed solution was then loaded onto a Mono-Q column (Qiagen). The Ulp1(403-621)p -Smt3(13-101)p complex was eluted from the column with 0.02-0.5 M NaCl gradient. Fractions were analyzed by native gel electrophoresis. Ulp1(403-621)p -Smt3(13-

101)p was purified further by size exclusion chromatography on a Superdex75 prep-grade column (Pharmacia). Fractions were pooled, concentrated to about 3.5 mg/mL, and stored in 20 mM Tris-HCl containing 50 mM NaCl at -80° C.

Crystallization and data collection

Crystals were grown at 21° C by the hanging drop vapor diffusion method. Ulp1(403-621)p -Smt3(13-101)p complex was mixed with an equal volume of reservoir buffer containing 0.1 M MES (pH=6.5), 10% (w/v) polyethylene glycol 20000, and 3% (w/v) 1,6-hexandiol. Crystals were grown over several days to a size of 0.05x0.1x0.3 mm. Crystals were stabilized in reservoir solution for heavy atoms soaks and pre-incubated in reservoir solution plus 17% ethylene glycol. X-ray data used to collect native and heavy atom sets were collected using crystals at a laboratory copper Kα source (Rigaku RU200) equipped with a confocal optics and a Raxis-IV imaging plate detector system. The high-resolution native data set and a four-wavelength MAD data set were collected at the National Synchrotron Light Source (Brookhaven, NY) at beamline X4A on an ADSC Quantum-4 detector. Data were processed using DENZO and SCALEPACK. Ulp1(403-621)p -Smt3(13-101)p were crystallized in space group P21212 (a=125.8 Å, b=53.4 Å, c=54.3 Å, α=β=γ=90°).

Structure determination and refinement

2.6 Å electron density maps for the Ulp1(403-621)p -Smt3(13-101)p complex were initially derived using multiple isomorphous replacement (MIR) methods and solvent flattening (DM). Mercury sites were identified using Patterson methods and placed into MLPHARE for phase refinement. An initial trace into these maps was accomplished using the program O and the model was not refined. A selenomethionyl-derived Ulp1 (403-625)p and native Smt3(11-97)p crystal was subjected to a four-wavelength MAD experiment (Hendrickson, *et al.*, EMBO J., 1991, 9:1665-1672). The 1.8 Å MAD data was input into the program SOLVE (Terwilliger and Berendzen, Acta Crystallogr, 1999, D55:849-861). Six selenium sites were located and used to generate 1.8 Å phases. Solvent flattening was performed using the program DM (Cowtan, Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, 1994, 31:34-38). The initial model derived from the MIR

electron density was placed into the MAD electron density maps for manual building. Refinement was accomplished using the programs REFMAC (Murshudov *et al.*, Acta. Crystallogr., 1997, A50:157-163) and wARP/ARP (Lamzin and Wilson, Acta. Crystallogr., 1993, D49:129-149) for automated building and placement of waters. The model underwent three rounds of refinement at 1.8 Å with manual rebuilding between each round. Refinement was extended to 1.6 Å for another round. The current model contains 735 waters and 303 amino acids (Rfree=25.3, R=17.9) and has excellent geometry; 92.9% of residues are in the most favored region of the Ramachandran plot with none occurring in disallowed regions. The model contained all 219 amino acid residues of Ulp1(403-621)p plus two additional N-terminal amino acids that correspond to residues left behind after thrombin cleavage of the N-terminally HIS-tagged protein. Ulp1p Arg422 did not have adequate density to model its side chain. The model for Smt3(13-98)p contains 79 of the possible 87 amino acid residues. Electron density for N-terminal Smt3 residues 12 through 19 could not be observed and were left out of the model.

Yeast plasmids and strains

Standard techniques were used to grow and maintain yeast strains (Guthrie and Fink, Methods Enzymol., 1991; 194). Yeast strain W303-1A (MATa *ura3-1 ade2-1 trp1-1 his3, -1115, leu2-3, -112 can1-100*) was used to generate a Ulp1 null strain by one-step gene replacement procedure (Rothstein, Methods Enzymol., 1991; 194:281-301). The resulting strain was termed Ulp1Δ strain W303ULP1 (MATa *ura3-1 ade2-1 trp1-1 his3, -1115, leu2-3, -112 can1-100, Ulp1::LEU2*, pSE360-ULP1). *In vivo* complementation analysis was conducted using plasmid shaffle technique (Boeke *et al.*, Methods Enzymol., 1987; 154,164). W303ULP1 was transformed with pSE358 plasmid containing either wildtype or mutant Ulp1 under Ulp1 natural promoter. Yeast cells were streaked onto agar plates that lack tryptophan and incubated at about 30° C until yeast colonies were visible. Colonies were then restreaked onto agar plates containing and 5-fluoro-orotic acid to induce the loss of pSE360-ULP1.

Glacatose inducible expression

Wildtype or deletion mutant Ulp1 fragments were amplified by PCR. The fragments were digested with EcoR1 and BamH1 restriction endonucleases and ligated into PYX133 (*CEN TRP1* GAL promoter; Ingenius). W303-1A was transformed with the plasmid. Yeast strains were grown on plates lacking tryptophan, to select for cells that contain the plasmid of interest, at about 30° C. Positive yeast colonies were then streaked onto SD(-Trp) agar plates, containing 2% (w/v) raffinose, with and without 0.01% (w/v) galactose. Cells were incubated at 30° C until a color change in the yeast colonies was observed.

Point mutations were introduced to subcloned fragments using QuikChange Site-Directed Mutagenesis Kit (Qiagen).

Results

Ulp1(403-621)p displayed full catalytic activity in cleavage reactions with human SUMO-1 and yeast Smt3 to produce their respective mature protein forms (data not shown). Ulp1(403-621)p also showed full catalytic activity in deconjugation reactions with Smt3-GFPuv and His6-ubiquitin-Smt3-HA (data not shown), suggesting that the C-terminal fragment of Ulp1 (Ulp1(403-621)p) is able to (i) deconjugate large proteins and (ii) produce the mature forms of Smt3 and SUMO.

The secondary structure of Ulp1(403-621)p includes 7α helices and 7β strands. Ulp1(403-621)p exhibits structural similarities to other papain-like cysteine proteases in the active site, which includes the central α helix, 3β strands, and the catalytic triad (Cys-His-Glu/Asp) (Figure 1).

The Ulp1(403-621)p-Smt3(13-98)p structure is currently refined to 1.6 angstroms and includes all 219 amino acid residues of Ulp1(403-621)p and 79 of the possible 87 amino acid residues of Smt3(13-98)p. Coordinates of the Ulp1(403-621)p-Smt3(13-98)p structure are shown in Table 1 (present before the claims). The interaction between Ulp1 and Smt3 is described below in the context of six conserved Ulp1 motifs.

The interactions of motif 1 of Ulp1 with Smt3 include side-chain to main-chain hydrogen bonding, Van der Waals (VDW) contacts, and one salt bridge between Arg438 of Ulp1 and the Glu94 of Smt3.

Motif 2 of Ulp1 is involved in the recognition of residues at the C terminus of Smt3, including VDW contacts with the Gly-Gly motif of Smt3 by Trp448. Motif 2 also contributes (i) 3 hydrogen bonds to the main-chain of Smt3 strand 5 and (ii) 2 acidic residues (Glu455 and Asp451) that participate in salt-bridging interactions with 2 Smt3 basic residues (Arg64 and Arg71).

Motif 3 of Ulp1 provides direct hydrogen bonds to the main-chain of Smt3, Gly69, by a conserved Asn472. Hydrogen bonds are also made between Ser473 and Thr477 to an invariant Smt3 residue Gln95 in strand 5. VDW contacts in the Smt3-Ulp1 interface is contributed by Phe474.

Motif 4 of Ulp1 contains salt bridging moieties (Arg489 and Arg493) that make contact with two acidic residues in Smt3 (Asp68 and Asp82). Positions 489 and 493 also produce hydrogen bonds through water to make interactions with Asp87 in Smt3. Motif 4 contributes VDW contacts in the Ulp1-Smt3 interface through Trp490.

Motif 5 of Ulp1 contains His514, which is the general base of the catalytic triad. Motif 5 also participates in several interactions with the C-terminal tail of Smt3. Motif 5 contains Asn509 and Gln512, which participate in hydrogen bonds with main-chain atoms of Smt3 residues 94 and 96. The carbonyl of Ser513 is involved in a main-chain to main-chain hydrogen bond with Smt3 Gly98.

Motif 6 of Ulp1 contains active site amino acid residues Cys580 and Gln574, which are involved in the function of the protein.

* * *

The present invention is not to be limited in scope by the specific embodiments described herein. Indeed, various modifications of the invention in addition to those described herein will become apparent to those skilled in the art from the foregoing description and the accompanying figures. Such modifications are intended to fall within the scope of the appended claims.

It is further to be understood that values are approximate, and are provided for description.

Patents, patent applications, publications, procedures, and the like are cited

throughout this application, the disclosures of which are incorporated herein by reference in their entireties.